UNIVERSITÄT
BERN

Faculty of Business, Economics
and Social Sciences

Department of Economics

Decomposing Differences in Arithmetic Means: A Doubly-Robust Estimation Approach

Boris Kaiser

13-08

October 2013

DISCUSSION PAPERS

Schanzeneckstrasse 1
Postfach 8573
CH-3001 Bern, Switzerland
http://www.vwi.unibe.ch

Decomposing Differences in Arithmetic Means: A Doubly-Robust Estimation Approach

Boris Kaiser*

October 25, 2013

Abstract

When decomposing differences in average economic outcome between two groups of individuals, it is common practice to base the analysis on logarithms if the dependent variable is nonnegative. This paper argues that this approach raises a number of undesired statistical and conceptual issues because decomposition terms have the interpretation of approximate percentage differences in geometric means. Instead, we suggest that the analysis should be based on the arithmetic means of the original dependent variable. We present a flexible parametric decomposition framework that can be used for all types of continuous (or count) nonnegative dependent variables. In particular, we derive a propensity-score-weighted estimator for the aggregate decomposition that is "doubly robust", that is, consistent under two separate sets of assumptions. A comparative Monte Carlo study illustrates that the proposed estimator performs well in a many situations. An application to the union wage gap in the United States finds that the importance of the unexplained union wage premium is much smaller than suggested by the standard log-wage decomposition.

Keywords: Oaxaca-Blinder; Decomposition Methods; Quasi-Maximum-Likelihood; Doubly Robust Estimation; Arithmetic and Geometric Means; Inverse Probability Weighting

JEL: C10; C50; C51; J31

Acknowledgements: I am grateful to three anonymous referees, Michael Gerfin, Blaise Melly, Stefan Boes, Kaspar Wüthrich, Steven Stillman, and various seminar participants for helpful comments.

*Department of Economics, Schanzeneckstrasse 1, University of Bern, CH-3001 Bern, Switzerland. E-mail: boris.kaiser@vwi.unibe.ch. Phone: +41 31 631 40 49.

1 Introduction

Decomposition methods are very useful tools to analyze differences in average economic outcomes between groups of individuals. Although a strand of the recent literature emphasizes the importance of decomposing the entire distribution (e.g. Machado and Mata, 2005; Melly, 2005), the mean differential remains an important summary statistic in decomposition analysis as it is simple ("one number"), easily understood and sometimes the direct object of interest of policy analysis (e.g. health care expenditures). Differences in means are usually decomposed with the Oaxaca-Blinder method (Oaxaca, 1973; Blinder, 1973) which splits the gap into a structural effect (due to differences in coefficients) and a composition effect (due to differences in covariates). In many empirical applications where outcomes are continuous and nonnegative, the dependent variable is log-transformed before performing the decomposition. While the log-transformation is usually done to allow for convenient estimation, it changes the interpretation of the outcome gap in an important way: in effect, the quantity to be decomposed when outcomes are in logs corresponds to a first-order approximation to the percentage difference in geometric means (GM). This quantity cannot be re-transformed to the first moments of the original dependent variable, the arithmetic means (AM).

This paper argues that a decomposition based on the original dependent variable may be preferable to a decomposition based on logarithms on statistical and conceptual grounds. Foremost, if the support includes zero, the GM is theoretically undefined. In applications, the dependent variable is sometimes artificially re-scaled in order to be able to take logarithms, but this distorts the distribution at the lower bound of the support (Santos Silva and Tenreyro, 2006). Furthermore, even if there are no zeros in the data, a decomposition of AMs can be preferable for several reasons. First, the log outcome gap is not invariant to changes in the higher-order moments of the distributions even if arithmetic means remain constant (Leslie and Murphy, 1997). If for example dispersion increases, the log outcome gap will change as well. This is clearly an undesired property for a measure of the difference in average outcomes. Second, the log outcome gap only offers an approximate interpretation, whereas the raw outcome gap offers an exact interpretation. Finally, the AM is the more common and intuitive summary statistic.

In this paper, we suggest modelling the untransformed outcome variable directly to decompose differences in arithmetic means. Important examples, in which decompositions of such outcome variables are of interest, include wages and earnings (Firpo et al., 2011, for an overview), health care expenditures (Vargas Bustamante and Chen, 2011), financial wealth (Barsky et al., 2002), firm-level productivity (Mueller, 2012), revenues (Munn and Hussain, 2010), students' test scores (Krieg and Storer, 2006), or counts, such as the number of cigarettes smoked (Bauer et al., 2007). Since in most of these examples, the conditional expectation function is thought to be convex, as reflected in the widespread

use of log-transformations, a linear outcome model is inappropriate (see e.g. Barsky et al., 2002). Instead, we propose a flexible nonlinear parametric framework based on an exponential regression model. An important difference to previous papers on nonlinear decompositions (Fairlie, 2005; Bauer and Sinning, 2008) is that we show how the counterfactual of interest can be identified by invoking the results from the treatment effects literature (Rubin, 1974).

As our methodological contribution to the literature, we suggest a new estimation strategy for decomposing differences in average outcomes when outcomes are nonnegative. We propose a "doubly robust" weighted Poisson quasi-maximum-likelihood (WPQML) estimator. The existence of this estimator is briefly mentioned in Wooldridge (2007), but to the best of our knowledge, it has not been studied in detail nor has it been used in empirical work. Quasi-maximum-likelihood (QML) estimators in general are attractive because they only require correct specification of the conditional mean function for consistency and no further distributional assumptions (Gourieroux et al., 1984). The Poisson QML estimator achieves some additional robustness if augmented with appropriate propensity-score weights. We show formally that the decomposition can be consistently estimated if *either* the outcome model *or* the propensity score model is correctly specified. Therefore, double robustness is a useful property to guard against misspecification.

To illustrate how the proposed estimator of the nonlinear decomposition performs in practice, we undertake a small Monte Carlo exercise. We compare the performance of weighted and unweighted QML estimators, linear and quadratic regression and reweighted regression (Firpo et al., 2007) under various scenarios with regard to functional form, overlap of covariate distributions and heteroskedasticity. We find that the weighted Poisson QML estimator produces convincing results in many situations. It should therefore prove to be an attractive estimation strategy for applied researchers who wish to perform decompositions of nonnegative outcomes.

In an empirical application to the union wage gap in the United States, we compare the approach of decomposing arithmetic means with the standard Oaxaca-Blinder decomposition based on the log-wage gap (geometric means). The former suggests that the union wage premium is much less important in explaining the wage differential between union and nonunion workers than the latter. The exact same finding emerges from a decomposition of the native-immigrant wage gap in Switzerland. In other words, the concept of the mean can have striking implications for the analysis of differences in economic outcomes and therefore requires more careful consideration than it usually receives in empirical research.

The remainder of this paper is organized as follows: Section 2 deals with the conceptual and statistical issues of decomposing differences in either AMs or GMs. In Section 3, we define the general decomposition framework and present the identification result for the counterfactual of interest. In Section 4, we discuss model specification and derive

the doubly robust estimator for the decomposition. Section 5 contains the Monte Carlo exercise and Section 6 the empirical applications. Section 7 briefly covers some extensions to detailed decompositions, endogeneity and sample selection and Section 8 contains some concluding remarks.

2 Arithmetic and Geometric Mean Differentials

2.1 Analytical Framework

The analytical framework used in this paper will be based on the potential-outcomes notation popularized by Rubin (1974). While questions of identification have mostly been ignored in the more traditional decomposition literature, the treatment effects framework clarifies the assumptions needed to identify the decomposition terms (Firpo et al., 2011). We begin by defining a large population indexed by $i = 1, 2, ..., N$ that contains two mutually exclusive and non-empty groups of individuals. Let $D_i = 1$ if person i belongs to group 1 and $D_i = 0$ if she belongs to group 0. The potential outcome of person i if she belongs to group g is y_{ig} , where $y_{ig} \in \mathbb{R}_0^+ \; \forall i$ and $g \in \{0, 1\}$. The potential outcomes and the realized outcome are linked by $y_i = D_i y_{i1} + (1 - D_i) y_{i0}$. Hence, we do not observe y_{i1} for units in group 0 and y_{i0} for units in group 1, which is why these quantities are "counterfactual". The notation introduced above will be used throughout this paper.

2.2 Measures of Mean Differentials

Before dealing with the technicalities of decomposition analysis, it is sensible to first clearly define the object under study, i.e. the quantity to be decomposed. In most situations we are interested (among other statistics) in the average gap in economic outcomes between two groups, a prominent example being the average wage gap between male and female workers. It seems natural to consider the difference in expected outcomes which is based on population **arithmetic means** (AM):

$$\Delta_{AM} = E[y_i|D_i = 1] - E[y_i|D_i = 0] \tag{1}$$

Of course, the outcome gap may also be expressed in terms of a percentage differential, e.g. $\%\Delta_{AM} = (E[y_i|D_i = 1] - E[y_i|D_i = 0])/E[y_i|D_i = 0]$, which requires the choice of a reference group (in this case group 0), but it is free from the underlying units in which outcomes are measured. Alternatively, if the decomposition is based on log outcomes, we have

$$\%\Delta_{GM}^{approx} \approx E[\ln y_i|D_i = 1] - E[\ln y_i|D_i = 0], \tag{2}$$

which is the approximate percentage differential in the population **geometric means** (GM). An exact percentage differential in terms of GMs is given by $\%\Delta_{GM} = [\exp(E[\ln y_i|D_i = 1]) - \exp(E[\ln y_i|D_i = 0])]/\exp(E[\ln y_i|D_i = 0])$ and the absolute differential in terms of GMs is $\Delta_{GM} = \exp(E[\ln y_i|D_i = 1]) - \exp(E[\ln y_i|D_i = 0])$. The approximate nature of (2) comes from a first-order Taylor expansion of $\%\Delta_{GM}$ around $E[\ln y_i|D_i = 1] - E[\ln y_i|D_i = 0] = 0$ that becomes more accurate as $\%\Delta_{GM}$ approaches zero.[1]

Is is notable that in many applications of nonnegative dependent variables, researchers prefer to decompose (2) instead of (1). In particular, this is often the case in studies of wages, where log-linear wage models are estimated by OLS to perform Oaxaca-Blinder-type decompositions. This is the dominant procedure because wages and covariates are usually assumed to have a log-link relationship, but also because it is standard practice to estimate linear models instead of nonlinear models. While the geometric-mean interpretation of (2) was noted by Oaxaca (1973), it is not explicitly mentioned in most empirical applications.

The question obviously arises as to how the concepts of AM and GM are related. Due to Jensen's inequality, we always have that $AM > GM$. However, the relationship between Δ_{AM} and Δ_{GM} is ambiguous without distributional assumptions and will depend on the properties of the distributions under study.

2.3 Statistical and Conceptual Issues

Left aside issues of modelling and estimation, which measure of the mean differential should be chosen when performing decompositions? There are several arguments to be made. Foremost, if the support of y_i includes zero, the GM is not defined. Thus, the decomposition of means must be based on AMs for outcomes that may be zero, such as individual health care expenditure, wealth, or count variables. Using ad hoc manipulations such as $\ln(1 + y_i)$ to accommodate zeros is inappropriate as it may severely distort the distribution if the fraction of zeros is non-negligible (cf. Santos Silva and Tenreyro, 2006).

Furthermore, even if the support of the dependent variable excludes zero (as is the case with wages) there are a number of arguments why a decomposition of means should be based on AMs instead GMs, or equivalently, on raw outcomes instead of log outcomes. First, as Leslie and Murphy (1997) note, the decomposition based on log outcomes is not invariant to changes in the higher-order moments of the distribution. As a result, even if the AMs of two distributions are the same, the difference in average log outcomes is

[1] Of course, the approximation in (2) could be made more exact if higher-order terms are included in the Taylor expansion. For simplicity, write $z_g \equiv E[\ln y_i|D_i = g]$ for $g = \{0, 1\}$, then $\%\Delta_{GM} = \exp(z_1 - z_0) - 1$. A Taylor expansion around $z_1 - z_0 = 0$ yields $(z_1 - z_0) + \frac{1}{2!}(z_1 - z_0)^2 + \frac{1}{3!}(z_1 - z_0)^3 + ...$

non-zero if, for example, the dispersion differs across distributions. We illustrate this by two simple examples:

Example 1. (Leslie and Murphy, 1997) Let there be two samples with only two observations each: $S^0 = \{2, 4\}$ and $S^1 = \{1, 5\}$. Then we have $\bar{y}^0 = \bar{y}^1 = 3$ and $\hat{\Delta}_{AM} = 0$, but the log gap equals $\%\hat{\Delta}_{GM}^{approx} = 0.47$.

Example 2. Suppose y^0 and y^1 are log-normally distributed with $E[y^0] = E[y^1] = 1$, but variances are unequal: $V[y^0] = 0.5 < V[y^1] = 1$. In this case $\Delta_{AM} = 0$, but there is a gap in log means given by $\%\Delta_{GM}^{approx} = E[\ln y_1] - E[\ln y_0] = (-.203) - (-.347) = .144$.[2]

As demonstrated above, if two distributions are identical up to a mean-preserving spread, the log-wage gap picks up the difference in dispersion. For a measure of the difference in means, this is a rather undesired property.

Second, the approximation in (2) becomes inaccurate for large differences in outcomes due to the concavity of the logarithm: for exact GM differentials of 5, 10, 20, and 50 percent, the "approximation bias" is 0.12, 0.47, 1.77 and 9.45 percentage points, respectively. This is clearly important in practice, since many differentials (e.g. male-female wage gap) are quite substantial in magnitude. In fact, the approximative nature of (2) is only seldom pointed out in empirical studies and sometimes no reference to the GM interpretation is made (e.g. Darity et al., 1995; Jürges, 2002). As a consequence, interpreting log-points as percentage differentials can be misleading.[3]

Third, apart from time-series contexts, it is customary to report arithmetic means rather than geometric means when summary statistics are presented. Although this does not represent a strong argument in favour of the AM decomposition per se, it nonetheless implies that the AM is usually the statistic of interest.

Finally, a potential weakness of the AM relative to the GM as a measure of central tendency is that it is more strongly affected by the presence of outliers in small samples.[4] Although the GM is indeed more robust, the argument only applies if the two distributions under study are adversely affected by outliers. This is because we are interested in *differences* in means and not in means *per se*.

The discussion above provides the motivation for developing a decomposition framework for differences in AMs that can be broadly applied to nonnegative dependent variables.

[2] Of course, examples of the converse case, where $\%\Delta_{GM} = 0$ and $\%\Delta_{AM} \neq 0$, can also be given.

[3] To avoid the problem described above, the log gap is often interpreted in terms of log-points or "log-dollars" (Black et al., 2006; Edin and Richardson, 2002; Kim, 2010). While this is formally correct and exact, talking about log-points arguably conveys little meaning in policy analysis.

[4] The most robust measure of central tendency is the median. It is unfortunate, however, that a direct extension of OB-type decompositions to the median wage gap is not possible since the law of iterated expectations does not hold for quantiles. See Firpo et al. (2011, Section 4) for a discussion of the various approaches to quantile decompositions.

3 Decomposition Framework and Identification

As for any decomposition problem, we must define a meaningful counterfactual. In the classic Oaxaca-Blinder decomposition of wages, the counterfactual is implicitly implied by the reference wage structure (Cotton, 1988; Neumark, 1988; Jann, 2008). We follow the approach of Firpo et al. (2011) and choose the "simple" counterfactual outcome $\mu^C \equiv E[y_{i0}|D_i = 1]$. This term corresponds to the mean outcome in group 1 that would prevail if their distribution were generated by the conditional expectation function (CEF) of group 0. Since the labelling of group 0 and 1 is arbitrary, the choice of the counterfactual largely depends on the question of interest. However, we find this particular counterfactual useful because it offers a simple interpretation and has a meaningful treatment-effect equivalent.[5] As opposed to the counterfactual, the group-specific means are directly observable, i.e. $E[y_i|D_i = g] = E[y_{ig}|D_i = g]$. As a result, the aggregate AM decomposition is

$$\Delta_{AM} = \underbrace{(E[y_i|D_i = 1] - E[y_{i0}|D_i = 1])}_{\equiv \Delta^S} + \underbrace{(E[y_{i0}|D_i = 1] - E[y_i|D_i = 0])}_{\equiv \Delta^X}, \tag{3}$$

The first term on the right-hand side is the **structural effect** (also: coefficients effect). In treatment effects terminology, it corresponds to the population average treatment effect on the treated (PATT), see e.g. Imbens and Wooldridge (2009). The second term is the **composition effect** (also: characteristics effect) and captures the part of the gap due to differences in covariate distributions across groups.

A prominent result from the treatment effects literature (see e.g. Imbens, 2004) states that identification of μ^C is ensured under the following set of assumptions:

$$y_{i0} \perp\!\!\!\perp D_i|X_i \tag{4}$$

$$p(X_i) < 1 \tag{5}$$

where $p(X_i) \equiv P(D_i = 1|X_i)$ is the propensity score and X_i is a row vector of covariates with support $\mathcal{X} \subseteq \mathbb{R}^k$. The conditions in (4) and (5) are referred to as the conditional independence assumption (CIA) and the common support assumption (CSA), respectively.

[5] In the traditional literature on OB wage decompositions, a number of alternative reference structures (i.e. counterfactuals) have been suggested (Jann, 2008), but these alternative measures sometimes imply undesired assumptions on the underlying structural effect. First, if the counterfactual is based on Neumark (1988), the coefficient estimates are likely to suffer from an omitted variable bias (see Jann, 2008). Second, if the counterfactual is based on the coefficients from the (population) regression in the pooled sample on the covariates and a dummy for the treated group, the wage structure effect is restricted to be the same for every individual (see Firpo et al., 2011). Third, if the counterfactual is a sample-size weighted average of the two wage structures as in Cotton (1988), the wage structure effect can be shown to equal $P(D_i = 0)PATT + P(D_i = 1)PATU$ under a set of baseline assumptions, with PATU being the population average treatment effect on the untreated (see Słoczyński, 2012). Clearly, this quantity differs from the population average treatment effect (PATE) unless $P(D_i = 0) = P(D_i = 1)$, and has therefore no interesting analogue in the treatment effects framework.

Henceforth, we will collectively refer to them as ignorability.[6] The former states that, once we control for X_i, the distribution of unobserved characteristics is the same across groups in the absence of treatment. The latter states that the support of the covariate distribution in group 1 (the treated) must be contained within the support of the covariate distribution in group 0 (the controls). Since these assumptions are well known in the literature, we refer to Imbens and Wooldridge (2009) for more discussion. Given these ignorability assumptions, identification of the counterfactual outcome can be stated as follows (see Imbens, 2004):

$$E[y_{i0}|D_i = 1] = E_X\{E[y_i|X_i, D_i = 0]|D_i = 1\} \tag{6}$$

To decompose differences in arithmetic means, the main issue that must be addressed is to specify and estimate an appropriate model for $E[y_i|X_i, D_i = 0]$, which we turn to in the next section.

4 Model Specification and Estimation

Since researchers are usually also interested in detailed decompositions, parametric models are more practical tools such that will not discuss nonparametric methods. For the ensuing discussion, we assume that the CEFs can be represented by $E[y_i|X_i, D_i = g] = \mu(X_i, \beta^g)$ for $g = \{0, 1\}$ and all $X_i \in \mathcal{X}$, where $\beta^g \in \mathbb{R}^k$ is a unique column vector of population coefficients.[7] As a useful starting point for modelling, consider the "generalized" functional form (Wooldridge, 1992):

$$\mu(X_i, \beta^g) = (1 + \lambda X_i \beta^g)^{1/\lambda} \text{ , for } g = \{0, 1\}, \tag{7}$$

which encompasses the linear regression model ($\lambda = 1$) and the exponential regression model ($\lambda \to 0$) as special cases.[8] We discuss these two specifications for modelling the dependent variable in the next two subsections.

4.1 A Note on the Linear Model

If the CEF of the outcome is assumed to be linear ($\lambda = 1$), then by the law of iterated expectations, the AM decomposition takes the familiar Oaxaca-Blinder form. Why should we not use a linear model that can be conveniently estimated with OLS? In the case of

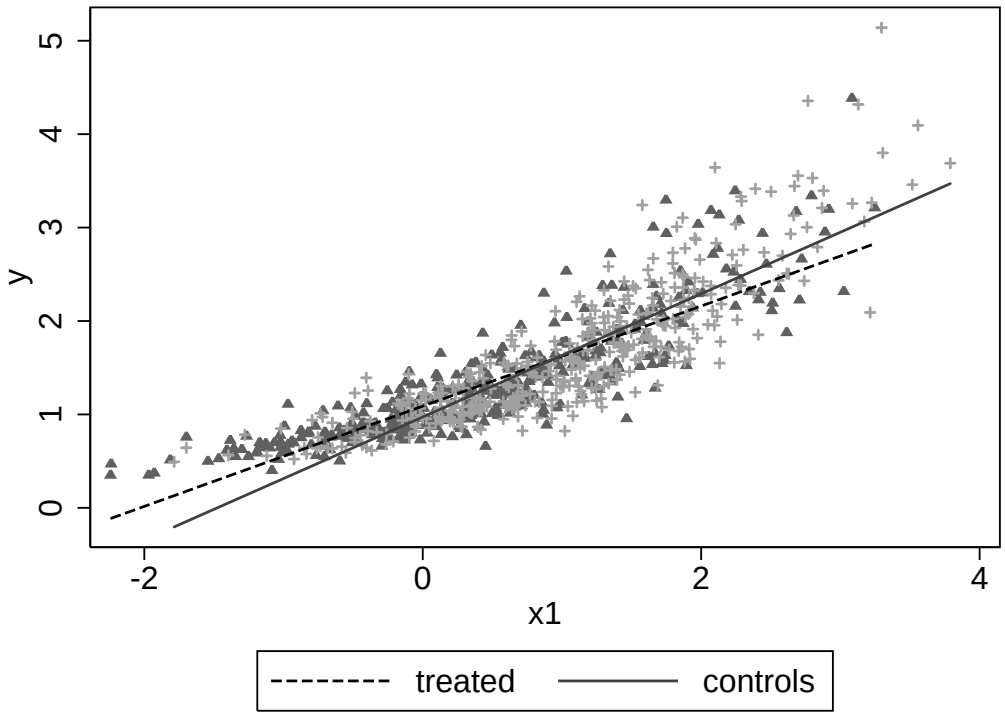
[6] Assumptions (4) and (5) are weaker versions of the assumptions necessary to identify the population average treatment effect (PATE).

[7] The uniqueness or identifiability of β^g rules out perfect multicollinearity among the covariates.

[8] In the language of generalized linear models, the first assumes an identity-link between outcomes and covariates and the second assumes a log-link relationship (equivalent to the functional form assumption of the log-linear model).

nonnegative dependent variables, the linear model appears unnatural because it does not restrict the support of the CEF in any way. Given nonnegative outcomes, the true CEF is likely to be nonlinear. While OLS offers the best *linear* approximation to the CEF, the approximation may still be very poor if the true CEF is, for example, highly convex (Barsky et al., 2002). To see this, consider the stylized example in Figure 1, where outcomes in both groups are generated by the same CEF ($\Delta^S = 0$), but the covariate distributions in the treated group and the control group are different, $N(0.5, 1)$ and $N(1, 1)$ respectively, which means ($\Delta^X \neq 0$). The two group-specific regressions of y_i on X_i yield

Figure 1: Linear Approximation



different parameters, meaning that the misspecified model leads to a spurious structural effect ($\hat{\Delta}^S \neq 0$). This inconsistency persists if ignorability holds. Moreover, Barsky et al. (2002) point out that the performance of the linear estimator becomes particularly poor if there is missing overlap in the covariate distributions due to the linear extrapolation of the misspecified CEF outside the common support.

There are several ways as to how the above described problems can be alleviated. First, one can try to improve the approximation to the CEF by adding higher-order terms of the covariates. In the case of high dimensional X_i, however, this may be cumbersome and lead to very noisy estimates. Furthermore, it complicates matters if one is interested in performing detailed decompositions. Second, another alternative is to use reweighted regression as suggested by Firpo et al. (2007). Their estimator is doubly robust (in the spirit of Robins et al., 2007) in that it consistently estimates the counterfactual if either the

CEF of the outcome model is linear or if the assumed propensity score model is correct. Intuitively, the propensity-score adjustment reduces the imbalance between the group-specific covariate distributions and thus reduces the contamination of the structural effect stemming from these imbalances. While reweighting generally increases the robustness of regression with respect to departures from linearity, it may be problematic to rely on the correct specification of the propensity score alone because any misspecification immediately leads to biased estimates.

4.2 Nonlinear Outcome Model

Due to the described weaknesses of the linear model in the case of nonlinear CEFs, a nonlinear specification that is consistent with the dependent variable being nonnegative for arbitrary values of β^g is generally preferable. This property is satisfied by the exponential model, which follows from (7) when $\lambda \to 0$. The model can be written as

$$y_{ig} = \exp(X_i\beta^g) + \varepsilon_{ig} \quad \text{for } g = \{0, 1\} \tag{8}$$

where ε_{ig} is an error term,[9] and the assumed CEF is $\mu(X_i, \beta^g) = \exp(X_i\beta^g)$.[10] The exponential CEF is used in many different models of nonnegative dependent variables. In the case of count data outcomes, it emerges from the Poisson model. In the case of continuous nonnegative outcomes, it follows from the gamma and exponential distributions. The exponential model is also very widely used in practice. In models of health care expenditure, for example, it is the standard functional form assumption (Mullahy, 2009). In wage models, the dependent variable used for estimation is usually in logarithmic form which implies the functional form assumed in (8). It is justified theoretically (Mincer, 1974) and empirically (Heckman and Polachek, 1974; Blackburn, 2007). Therefore, the exponential model seems a suitable choice for many contexts.

[9] In (8), we may use a multiplicative or additive error term because we will only need to impose mean-independence (Wooldridge, 1992).

[10] Note that for the aggregate decomposition, regressors may be correlated with unobservables, as long as the conditional distribution of these unobservables is the same across group (i.e. as long as ignorability holds). Thus, β^g may have a "reduced-form" interpretation if we are only interested in the aggregate decomposition. For the detailed decomposition, however, regressors must be exogenous such that parameters have a "structural" interpretation.

4.2.1 Aggregate Decomposition

If the outcome model is (8), the law of iterated expectations implies that the aggregate decomposition defined in (3) can be written as

$$\begin{aligned}\Delta_{AM} = &\underbrace{E[\exp(X_i\beta^1)|D_i=1] - E[\exp(X_i\beta^0)|D_i=1]}_{\Delta^S} \\ &+ \underbrace{E[\exp(X_i\beta^0)|D_i=1] - E[\exp(X_i\beta^0)|D_i=0]}_{\Delta^X}\end{aligned} \tag{9}$$

The only critical expression is the counterfactual, $\mu^C = E[\exp(X_i\beta^0)|D_i = 1]$, which is identified if ignorability holds.

4.2.2 Doubly Robust Estimation

To estimate the decomposition in (9), expectations are replaced by sample analogues and population coefficients by $\hat{\beta}^1$ and $\hat{\beta}^0$, which are estimated coefficients from the two group-specific samples. Estimation can be based on quasi-maximum likelihood (QML) techniques, also referred to as estimation of generalized linear models (GLM) in the statistics literature. It is a well known result by Gourieroux et al. (1984) that QML estimators are consistent *regardless* of the underlying distribution of the data, as long as the CEF is correctly specified. In other words, these models can be estimated consistently with minimal distributional assumptions. The difference between the various QML estimators is simply how observations are weighted in the first-order conditions of the maximization problem, leading to different requirements for asymptotic efficiency. Several authors (Manning and Mullahy, 2001; Santos Silva and Tenreyro, 2006) find that Poisson and gamma QML estimators perform well in many situations, while Gaussian QML (nonlinear least squares) leads to more erratic results, especially in heteroskedastic environments.

A major advantage of the QML framework with regard to the decomposition in (9) is that a so-called "doubly robust" estimator for the counterfactual mean can be derived. Double robustness refers to an estimator that is consistent if either the CEF of the outcome model or the propensity score model (or both) is correctly specified. As Wooldridge (2007) briefly mentions in his discussion on average treatment effects, Poisson QML can be made doubly robust if augmented with the appropriate propensity-score weighting function. We will refer to this estimator as the WPQML estimator. Denote the probability limit of the WPQML estimator by $\beta^{g*} = plim(\hat{\beta}^g)$, where we allow for the possibility that the CEF may be misspecified, i.e. $E[y_{ig}|X_i, D_i = g] \neq \exp(X_i\beta^{g*})$ for some $X_i \in \mathcal{X}$. However, the defining property of the WPQML estimator is that $E[y_{ig}|D_i = g] = E[\exp(X_i\beta^{g*})|D_i = g]$ will always hold by construction. As shown below, this property arises from the first-order conditions associated with this estimator. For the sub-population of interest (group 0), the WPQML estimator solves the sample analogue of the following population maximization

problem:

$$\max_{b^0} E[\omega(X_i)(y_i X_i b^0 - \exp(X_i b^0) - y_i!)|D_i = 0],$$

where $\omega(X_i)$ is the propensity-score weighting function defined below. The corresponding population first-order conditions (FOCs) are:

$$E[\omega(X_i)(y_i - \exp(X_i \beta^{0*}))X_i'|D_i = 0] = 0. \tag{10}$$

Since Wooldridge (2007) does not touch upon the WPQML and confines the discussion to the population average treatment effect (PATE), Theorem 1 below shows that WPQML with appropriate weights yields a doubly robust estimator for our counterfactual of interest, and thus also for the structural effect Δ^S and the characteristics effect Δ^X of the decomposition.

Theorem 1. *Define the weighting function by* $\omega(X_i) \equiv \frac{\Lambda(X_i)}{1-\Lambda(X_i)} \frac{1-p}{p}$, *where* $\Lambda(X_i)$ *is a model for* $p(X_i) \equiv P(D_i = 1|X_i)$ *and* $p \equiv P(D_i = 1) > 0$. *Assume that (i) we have a random sample from an i.i.d. population (ii) all relevant moments of* y_i *and* X_i *exist and are finite and (iii) ignorability (assumptions* (4) *and* (5)*) holds.*

Then, the WPQML estimator is doubly robust in the sense that it identifies the counterfactual of interest through $E[y_{i0}|D_i = 1] = E[\exp(X_i \beta^{0*})|D_i = 1]$, *if either the outcome model or the propensity score model is correctly specified (or both, of course). The counterfactual mean can be consistently estimated by the appropriate sample analogue* $\frac{1}{N_1} \sum_{i:D_i=1}^{N_1} \exp(X_i \hat{\beta}^0_{WPQML})$.

The proof is given in Appendix A.

The central implication of Proposition 1 is the second part of the double robustness property (shown in Part III of the proof): the counterfactual is identified even if the WPQML estimator does not converge to the parameter vector of the true CEF ($\beta^{0*} \neq \beta^0$) provided that the propensity score model is correctly specified. This is the main idea of the double robustness property because identification and consistent estimation are ensured under two separate sets of assumptions. Therefore, double robustness guards better against misspecification than relying on correct specification of the outcome model alone or the propensity score model alone. Although the double robustness property is very appealing, it is important to remember that the result does not automatically extend to the detailed decomposition; if we fail to identify the parameters of the CEF, a detailed decomposition into the contributions of individual covariates to the gap is no longer guaranteed to be consistent.

Due to the nonlinear model and the two-step nature of the estimator, bootstrapping the entire procedure is the more practical alternative to conduct inference than deriving analytical standard errors via nonlinear GMM. Note that propensity score weighting may also come at the cost of an efficiency loss if the CEF is correctly specified, but our Monte

Carlo exercise suggests that the opposite can also happen in the case of heteroskedasticity where the conditional variance of the error is proportional to the conditional mean.

5 Monte-Carlo Simulation

To assess and compare the various estimators of the proposed decomposition, we conduct a small simulation study. Specifically, the objective is to test how various estimation methods perform in estimating the decomposition under different data-generating CEFs of the outcome model, different scenarios for the common support of the covariate distributions, as well as heteroskedasticity in the outcome model.

5.1 Set-Up

The simulation design mimics a potential-outcomes framework in which the CEF of group 0 has larger coefficients than the CEF of group 1 (negative structural effect) and group 0 has higher average values of the covariates than group 1 (negative composition effect). We generate the covariates $X_1, X_2 \sim N(0,1)$ with $Corr(X_1, X_2) = 0.2$. The assignment of observations to the two groups is based on the following latent variable model:

$$\begin{aligned} D^* &= \delta_1 X_1 + \delta_2 X_2 + \xi, \qquad \text{with } \xi \sim N(0,1) \text{ and } \delta_1, \delta_2 > 0 \\ D &= 1[D^* < \text{Med}(D^*)] \end{aligned} \tag{11}$$

The threshold in the indicator function is equal to the median of the latent variable such that the two groups are of equal size. The group indicator variable is $D = 1$ for group 1 (the treatment group) and $D = 0$ for group 0 (the control group). The random assignment error ξ is stochastically independent of the processes determining outcomes, which means that the CIA is satisfied in our set-up. The parameters (δ_1, δ_2) determine how strongly group membership is correlated with covariates, in other words, how strongly the group-specific covariate distributions differ from one another. We consider two specifications for the latent-variable equation. In the first one, there is perfect common support. In the second one, common support is small and the overlapping support assumption is violated for a subset of the population, see Appendix B for more details.

To illustrate the consequences of departures from the standard log-link assumption of exponential models, we consider two separate cases for the functional form. Starting from the generalized functional form, $E[y_g|X] = (1 + \lambda X\beta^g)^{1/\lambda}$ for $g = \{0,1\}$, in the first case (the baseline specification) we assume $\lambda \to 0$ which reduces to the exponential model. In the second case, we set $\lambda = 0.5$. This is a quadratic specification and thus also ensures that outcomes are always nonnegative. The curvature of this specification

lies in-between the linear and exponential model.[11] It is important to note that we do *not* consider the linear model as a possible data generating process because, unlike the two functional forms above, it does not restrict the support of the dependent variable to be nonnegative. Since the outcome distribution is affected by the functional form, we choose different values for β^g such that mean outcomes are of comparable size. See Appendix B for more details on the parameterization of the CEFs.

Finally, the potential outcomes are generated by multiplying the CEF with an error term that is drawn from the log-normal distribution with $E[\varepsilon_g|x] = E[\varepsilon_g] = 1$. To test whether the estimators are sensitive to heteroskedasticity, we consider two cases for the conditional variance of the error term. The first is the benchmark case of homoskedasticity and the second is heteroskedastic, where the conditional variance of the error is positively related to the CEF, see Appendix B for more details.

The quantity associated with uncertainty is the counterfactual mean, which is why the focus lies on estimating $\mu^C = E[y_0|D = 1]$ in this Monte Carlo exercise. The goal is to have an estimator based on the observed outcomes that is close as possible to the estimates from the "true" estimator $\hat{\mu}^C_{true} = \frac{1}{N_1} \sum_{i:D=1}^{N_1} y_{i0}$. We compare the performance of the following estimators for μ^C:

i. WPQML: The estimation method is Poisson QML augmented with propensity-score weights as described in section 4.2.2. The propensity score is estimated by a logit model.

ii. PQML: unweighted counterpart of WPQML.

iii. GQML: The estimation method is gamma QML, see e.g. Blackburn (2008).

iv. OLS1: linear regression. The CEF is assumed to be linear in X_i.

v. OLS2: quadratic regression. The vector of covariates includes a second-order polynomial of X_i.

vi. WLS: reweighted linear regression estimator proposed by Firpo et al. (2007). The propensity score is estimated by a logit model.

We generate samples of size 2,000 such that the estimation of the counterfactual is based on 1,000 observations. The entire procedure is repeated 10,000 times.

5.2 Results

The results are presented in Table 1. To assess both bias and precision, we report the mean bias in percentage terms relative to the true estimates, as well as the root mean

[11] Of course, one could also assume that λ be group-specific. But we do not do that for tractability of the results.

squared deviation from the true estimates (referred to as RMSD). For better readability, RMSD is expressed relative to the RMSD of our benchmark, the WPQML estimator.

Table 1: Monte Carlo Results for Aggregate Decomposition

Panel A. Exponential CEF

		Perfect Common Support				Small Common Support			
		Hom.		Het.		Hom.		Het.	
		BIAS	RMSD	BIAS	RMSD	BIAS	RMSD	BIAS	RMSD
i.	WPQML	0.0002	1.000	-0.0080	1.000	-0.0283	1.000	-0.0185	1.000
ii.	PQML	0.0171	1.010	0.0424	1.007	0.0745	0.953	0.5203	1.774
iii.	GQML	0.0011	1.000	-0.0198	0.997	-0.0040	0.817	0.1905	1.054
iv.	OLS1	-5.2929	1.120	-5.2543	1.106	-63.7008	2.501	-63.2967	5.721
v.	OLS2	0.6296	1.047	0.6096	1.088	24.1469	3.817	23.1846	10.298
vi.	WLS	-0.0548	1.064	-0.0282	1.039	-3.8622	1.181	-3.8381	1.218

Panel B. Quadratic CEF

		Perfect Common Support				Small Common Support			
		Hom.		Het.		Hom.		Het.	
		BIAS	RMSD	BIAS	RMSD	BIAS	RMSD	BIAS	RMSD
i.	WPQML	-0.2862	1.000	-0.2066	1.000	-24.0096	1.000	-21.0950	1.000
ii.	PQML	-11.5516	0.785	-11.2407	0.827	-53.5176	0.281	-53.0108	0.085
iii.	GQML	-1.2452	0.941	-1.4599	0.905	-46.0314	0.363	-46.1981	0.094
iv.	OLS1	-18.9098	1.047	-18.7359	0.957	-197.1998	0.900	-196.6338	0.345
v.	OLS2	0.0591	0.650	0.2124	0.901	0.1895	1.019	-0.4942	0.565
vi.	WLS	-0.3566	1.025	-0.2030	1.054	-24.9757	0.989	-25.2470	0.260

Notes: BIAS refers to the mean percentage deviation from the experimental estimate multiplied by 100. RMSD refers to the root mean squared deviation from the experimental estimate and is normalized by the RMSD of the WPQML estimator for ease of comparison. Results are based on 10,000 iterations.

We first turn to the results from the exponential specification in Panel A of Table 1. Under perfect common support, performance of the QML estimators and WLS is generally good, but WPQML stands out with the smallest bias. Only OLS1 is clearly biased, which demonstrates that the best linear approximation to the CEF performs poorly in estimating μ^C when the true CEF is nonlinear. In contrast, OLS2 has both smaller bias, since the higher-order terms of the covariates capture the non-linearities. These findings emphasize the importance of correctly specifying the CEF. Moreover, the reweighted regression (WLS) considerably improves upon unweighted linear regression (OLS1). Overall, all estimators, except OLS1 and OLS2, perform quite well irrespective of the heteroskedasticity in the data.

If the common support is reduced, we notice a number of changes. First, the performance of OLS1 (and to a lesser extent OLS2) deteriorates considerably. This result is not surprising, since the estimator linearly extrapolates the misspecified CEF to the covariate space of missing overlap. Second, WLS is now biased and imprecise relative to

the QML estimators. Third, the WPQML estimator becomes more precise than PQML (and GQML) under heteroskedasticity because observations with large conditional mean are both more noisy and more likely to be outside the common support. In other words, propensity-score weights tend to assign less weight to noisy observations, thus increasing precision. This result is interesting because heteroskedasticity of the form modelled here is often present in real data.

Panel B reports the results where the CEF takes the quadratic form such that the QML estimators are now misspecified and OLS2 is correctly specified. WPQML (and WLS) continues to perform well alongside OLS2 as long as common support is satisfied. It is only after the overlapping support assumption is violated (small common support) that the weighted estimators also become biased because ignorability no longer holds here.

We summarize the main findings from this simulation study as follows. First, as expected, linear OLS1 performs poorly given a nonlinear CEF. Second, it can be improved upon if higher-order terms are included but this comes at the expense of more noisy estimates in finite samples. Third, the propensity-score weighted estimators both produce good results and are usually better than their unweighted counterparts. Finally, comparing results across various scenarios, we find that WPQML produces the most convincing results, as long as the common support assumption is satisfied.

6 Empirical Application

6.1 Union Wage Gap in the United States

We apply the decomposition methods to the union wage gap in the United States. The dataset is constructed from the Current Population Survey (CPS) in the year 2012 and contains workers aged between 18 and 64. We define nonunion workers to be the reference group or "controls" ($D_i = 0$) and union workers to be the "treatment group" ($D_i = 1$). The dependent variable of interest is hourly wages. Therefore, the counterfactual mean of interest is the average hourly wage union workers would earn if they were paid according to the wage structure of nonunion workers. The set of explanatory variables includes years of education, years of experience (quadratic), dummies for ethnicity, region, marital status, immigration and gender.

The descriptive statistics in Table 2 show a considerable average wage differential between the two groups. Union workers earn 4 dollars more than nonunion workers as measured in terms of AM. The log-wage gap is 0.23 and the corresponding GM wage gap in levels is 4.5 dollars and thus quite smaller than the AM wage gap. Comparing covariates, we see that union workers have more experience and are less likely to live in Southern states but are more likely to be married. For the other explanatory variables, differences seem more modest.

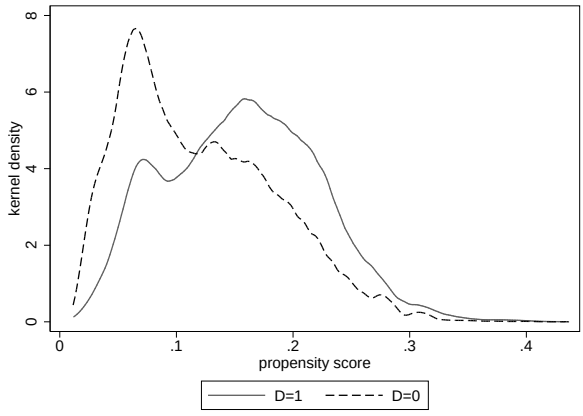
Table 2: Descriptive Statistics

	nonunion mean (st.dev.)	union mean (st.dev.)	difference
arithmetic mean wage	21.00	24.97	3.97
	(15.95)	(20.26)	
log of arithmetic mean wage	2.843	3.075	0.232
	(0.658)	(0.559)	
geometric mean wage	17.16	21.64	4.48
	(0.658)	(0.559)	
experience	21.15	24.50	3.35
	(12.69)	(11.72)	
years of education	13.73	14.22	0.49
	(2.61)	(2.49)	
black	0.116	0.141	0.025
other ethnic minority	0.089	0.074	−0.014
Midwest U.S.	0.220	0.242	0.022
Southern U.S.	0.392	0.189	−0.203
Western U.S.	0.215	0.285	0.069
married	0.543	0.624	0.081
divorced or widowed	0.142	0.156	0.014
foreign-born	0.181	0.142	−0.039
female	0.492	0.454	−0.038
# observations	128166	17330	

Notes: Standard deviations are in parentheses. Sampling weights are used. *Source*: Current Population Survey

An important lesson from our Monte Carlo exercise is that the decomposition is only sensible for those observations in the common support. We therefore inspect the overlap of the covariate distributions across the two groups. A straightforward way of doing this is to compare propensity score distributions across groups. Figure 2 shows kernel density

Figure 2: Density Estimate of the Propensity Score



estimates where the propensity score has been estimated with a logit model using the same set of covariates. We see immediately that estimated propensity scores are bounded well away from unity such that there are no observations in the data that violate the common support assumption in (5).

Table 3 presents the results of the decomposition of average wage differentials. We do

Table 3: Decomposition of the Union Wage Gap in the U.S. (2012)

Arithmetic mean (AM) decomposition (dependent variable: wage)

	$\hat{\Delta}$	(SE)	$\hat{\Delta}^X$	(SE)	%	$\hat{\Delta}^S$	(SE)	%
i. WPQML	3.977	(0.185)	2.604	(0.080)	65.49	1.372	(0.184)	34.51
ii. PQML	3.967	(0.169)	2.601	(0.080)	65.56	1.366	(0.162)	34.44
iii. GQML	4.061	(0.146)	2.486	(0.068)	61.23	1.574	(0.145)	38.77
iv. OLS1	3.967	(0.172)	2.747	(0.068)	69.23	1.221	(0.166)	30.77
v. OLS2	3.967	(0.164)	2.764	(0.077)	69.67	1.203	(0.155)	30.33
vi. WLS	4.093	(0.162)	3.088	(0.084)	75.45	1.005	(0.158)	24.55

Geometric mean (GM) decomposition (dependent variable: log-wage)

	$\hat{\Delta}$	(SE)	$\hat{\Delta}^X$	(SE)	%	$\hat{\Delta}^S$	(SE)	%
OLS ($\%\hat{\Delta}_{GM,approx}$)	0.232	(0.005)	0.129	(0.003)	55.46	0.103	(0.005)	44.54
OLS ($\hat{\Delta}_{GM}$)	4.483	(0.109)	2.357	(0.064)	52.58	2.126	(0.103)	47.42

Notes: Bootstrap standard errors (1000 iterations) in parentheses.

not compute detailed decompositions in order to focus on the essential task of estimating the aggregate decomposition. The top panel shows AM decompositions using various estimators and the bottom panels shows GM decompositions, i.e. decompositions of the log-wage gap. The QML estimates suggest that about 61-65% of the AM gap is due to different characteristics ($\hat{\Delta}^X$). The remainder is due to different wage structures ($\hat{\Delta}^S$) and often referred to as the average "union wage premium". If we regard union status as a treatment, $\hat{\Delta}^S$ can also be interpreted as the average treatment effect on the treated (ATT) (Firpo et al., 2011). In the data studied here, the OLS estimators (especially WLS) yield smaller union wage premiums than the QML estimators. We therefore argue that the choice of the estimator is important. The fact that all QML estimates are very similar lends support to the notion that the exponential model provides a good description of the CEF.

For the GM decomposition based on log-wage regressions, we report results both in logs ($\%\hat{\Delta}_{GM,approx}$) and in levels ($\hat{\Delta}_{GM}$). The former corresponds to the standard Oaxaca-Blinder decomposition of log-wages and the latter is obtained by exponentiating means measured in logs. Of course, the absolute values of AM and GM decompositions cannot be compared directly due to the different concepts of the mean. However, we can compare the shares of the composition effect and the structural effect in the total differential. The GM decomposition suggests that the union wage premium explains about 47% of the

average wage gap, while the AM decomposition suggests that it explains only about 35% (WPQML estimate). We therefore reach quite different conclusions when we base the decomposition on either AMs or GMs.

It is important to note that the discrepancy does not necessarily mean that one of the two decompositions is inconsistently estimated. It may entirely come from the fact that we apply the decomposition to *different distributions*, with one being a nonlinear (monotone) transformation of the other. Unfortunately, the size of the discrepancy depends on the properties of the distribution and is hard to investigate further. However, it may be instructive to point out that the configuration above, where the relative size of the union wage premium is larger in the log-wage decomposition, for example, also arises in the case where outcomes are log-normally distributed and the inequalities $E[\ln y_i|D_i = 0] < E[\ln y_{i0}|D_i = 1] < E[\ln y_i|D_i = 1]$ and $V[\ln y_{i1}] < V[\ln y_{i0}]$ hold.

Thus, choosing between the two concepts of the mean can matter in practice because we reach different conclusions when we evaluate the relative importance of the union wage premium in explaining the average wage differential.

6.2 Native-Immigrant Wage Gap in Switzerland

This section applies the various estimators of the AM decomposition introduced above to the native-immigrant wage gap in Switzerland. The data is drawn from the Swiss Earnings Structure Survey 2008. There are two main advantages of this dataset. First, the data are generally considered to be of higher quality than comparable survey data, since it is elicited directly from employers' records. Second, the large size of the dataset ensures that sampling bias is not a concern.

For the purpose of illustration, we confine the analysis to male full-time workers in German-speaking Switzerland.[12] In addition, individuals with missing information on education, foreigners with unknown residence status, and those aged under 20 or above 65 are excluded. These sample selection criteria still leave us with a very large sample size of more than 400,000 observations. The dependent variable of interest is full-time equivalent gross monthly earnings.[13] The following set of controls is used: educational attainment (9 categories), potential work experience (quadratic), tenure (quadratic), skill requirement level (4 categories) and marital status (3 categories). Swiss and foreign workers are taken to be group 0 and group 1, respectively. That is, $D_i = 1$ if person i has no Swiss citizenship and $D_i = 0$ if person i is native. As a result, we regard immigrants as the observational analogue of the treatment group and natives as the control group. As a consequence, the counterfactual outcome of interest is the mean wage of immigrants that would prevail if they were paid according to the wage structure of natives.

[12] "Full-time" are those who work at least 90% of a full-time equivalent.

[13] Including monetary benefits, extra pay for night-shifts or weekend-shifts and, where applicable, one twelfth of the 13th monthly salary

Table 4 summarizes descriptive statistics for wages and covariates across native and foreign workers. As we can see, immigrants constitute about a quarter of adult male

Table 4: Descriptive Statistics

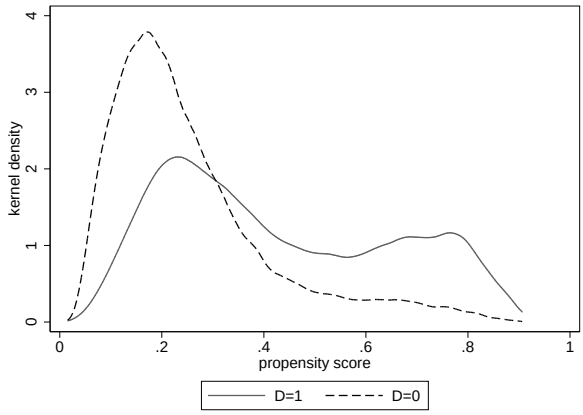
	natives mean (st.dev.)	immigrants mean (st.dev.)	difference
arithmetic mean wage (CHF)	7277	6422	−855
	(4845)	(4950)	
log wage	8.788	8.652	−0.136
	(0.411)	(0.420)	
geometric mean wage (CHF)	6557	5724	−834
	(0.411)	(0.420)	
work experience (years)	22.467	20.946	−1.521
	(11.733)	(10.513)	
tenure (years)	9.772	6.982	−2.790
	(9.852)	(7.902)	
education level			
university	0.058	0.083	0.026
college	0.069	0.046	−0.024
higher vocational training	0.149	0.065	−0.084
teaching diploma	0.003	0.002	−0.001
secondary school	0.016	0.013	−0.002
vocational training	0.619	0.419	−0.200
firm-specific vocational training	0.020	0.071	0.051
primary school	0.047	0.209	0.162
other education	0.021	0.093	0.072
skill requirement level			
very high	0.133	0.076	−0.057
high	0.380	0.256	−0.123
medium	0.405	0.409	0.004
low	0.082	0.258	0.176
marital status			
single	0.380	0.309	−0.071
married	0.540	0.625	0.084
divorced, widowed	0.080	0.067	−0.013
# observations	315192	134293	

Notes: The sample consists of male full-time workers from German-speaking Switzerland in 2008. Standard deviations are in parentheses. Sampling weights are used. *Source*: Swiss Wage Structure Survey, Swiss Federal Statistical Office.

employment. While AM and GM wage *differentials* are similar, AM and GM *wages* differ considerably across groups.

Before performing decompositions, it makes sense to investigate first the common

Figure 3: Density Estimate of the Propensity Score



support of the group-specific covariate distributions. We estimate the propensity score with a binary logit model using the same specification as in the outcome model. Kernel density estimates of the estimated propensity scores are depicted in Figure 3. We see that $\hat{p}(x_i)$ is bounded away from unity as there are no observations with values exceeding 0.9. Since only the upper threshold is relevant when estimating the counterfactual, we conclude that the common support assumption in (5) is satisfied.

The results of the mean wage decomposition between native and immigrant workers are shown in Table 5. Overall, the estimates of the AM wage decomposition suggest that

Table 5: Decomposition of the Wage Gap

Arithmetic mean (AM) decomposition (dependent variable: wage)

	$\hat{\Delta}$	(SE)	$\hat{\Delta}^X$	(SE)	%	$\hat{\Delta}^S$	(SE)	%
i. WPQML	-846.12	(22.39)	-667.43	(15.84)	78.88	-178.69	(17.49)	21.12
ii. PQML	-852.91	(20.66)	-624.12	(14.97)	73.18	-228.79	(16.89)	26.82
iii. GQML	-849.97	(19.69)	-622.63	(13.41)	73.25	-227.34	(15.60)	26.75
iv. OLS1	-852.91	(20.82)	-587.07	(15.25)	68.83	-265.83	(17.35)	31.17
v. OLS2	-852.91	(20.53)	-665.96	(15.02)	78.08	-186.95	(17.10)	21.92
vi. WLS	-860.20	(21.68)	-660.56	(15.51)	76.79	-199.64	(16.64)	23.21

Geometric mean (GM) decomposition (dependent variable: log-wage)

	$\hat{\Delta}$	(SE)	$\hat{\Delta}^X$	(SE)	%	$\hat{\Delta}^S$	(SE)	%
OLS ($\%\hat{\Delta}_{GM,approx}$)	-0.136	(0.002)	-0.087	(0.002)	63.942	-0.049	(0.002)	36.058
OLS ($\hat{\Delta}_{GM}$)	-831.52	(11.56)	-544.60	(9.53)	65.49	-286.92	(8.86)	34.51

Notes: Bootstrap standard errors (1000 iterations) in parentheses.

the largest part of the gap (about 70-80%) is explained by differences in characteristics (Δ^X) between natives and immigrants. The remaining part is the amount that the average immigrant earns less in monthly wages than the average native due to differences in the wage structure (Δ^S), which can be caused, for instance, by different returns to human

capital and/or discrimination. Comparing estimates, we note that the results from OLS1 are quite different from the rest. Generally, OLS2 and WLS are considered more reliable than OLS1 because the former leads to a better approximation of the unknown CEF and the latter achieves some additional robustness through weighting. As we can see, the WLS estimate is also closer to the WPQML estimate.

For the GM decomposition based on log-wage regressions, we report results both in logs and levels. Comparing results across the two concepts of mean decompositions, we notice an important quantitative difference: the GM estimates imply a considerably larger wage structure effect relative to the total gap. In fact, the wage structure effect in the log-wage decomposition accounts for some 36% of the total gap, which exceeds the WPQML estimate (21%) by about 15 percentage points. Only 1.4 percentage points of the discrepancy can be attributed to the approximation bias of the log-specification.

To summarize, the analysis on the immigrant-native wage gap in Switzerland produces the same qualitative findings as the analysis of the union wage gap in the United States. This reinforces the notion that the measure of the mean crucially affects the conclusions we draw with respect to the shares of Δ^X and Δ^S in the total wage gap.

7 Extentions

We briefly discuss how detailed decompositions can be performed in our framework and how issues of endogenous regressors (that violate ignorability) and sample selection can be addressed.

7.1 Detailed Decomposition

An important issue in decomposition analysis is to measure how much individual covariates contribute to the differential in mean outcomes. For example, in wage decompositions, it is interesting to examine how much differences in, say, education contribute to the wage gap. This type of question can be answered by performing detailed decompositions.

Generally, a detailed decomposition of the structural effect is difficult *regardless* of the dependent variable and the specified outcome model. The reason is that the the separate contributions of variables without a natural zero point are not interpretable (see Oaxaca and Ransom, 1999).[14] For this reason, detailed decompositions mainly focus on the composition effect.

In general, the detailed decomposition is less straightforward in nonlinear models than in linear models because the individual contributions of covariates are not additively

[14] Yun (2005) proposes a solution to the problem consisting of an ex-post normalization on the coefficients. However, as Firpo et al. (2011) note, such normalizations clearly come at the cost of interpretability.

separable. However, several approaches are available to perform detailed decompositions in nonlinear models, as the one discussed in Section 4.2. First, Yun (2004) uses a first-order Taylor approximation around the means of the covariates. The advantage is that the decomposition is path independent and satisfies the adding-up property (see Firpo et al., 2011, Chapter 2.2). The drawback is that it does not take into account that differences in higher-order moments (e.g. variances) affect outcomes through the nonlinearity of the CEF. Kaiser (2013) offers a method which is similar in spirit but takes into account the effects of such differences in covariate distributions. Second, Rothe (2012) offers a novel approach based on specifying copula functions to decompose the composition effect. The advantage of this method is that it can be applied to any type of outcome model included the one discussed in this paper. However, the flexibility of the approach comes at the expense of a more complex interpretation of the detailed decomposition terms.

7.2 Endogeneity of the Covariates

As discussed in Section 3, the structural effect and the composition effect are identified even if some of the regressors in X_i are correlated with unobservables, provided that ignorability still holds. (In contrast, the individual contributions of endogenous covariates are not identified in a detailed decomposition.)

Now consider the case where ignorability is violated due to endogenous covariates. In the exponential model, this problem can generally be addressed with instrumental variables as in the linear model. If an appropriate instrument is available, the parameters of the exponential models in (9) can be consistently estimated by the two-step control function approach described in Wooldridge (2010, Ch. 18), given that the CEF of the outcome model is correctly specified. Consider the case of a single endogenous variable, y_2, and instrument(s), Z_i, satisfying the usual exclusion restriction. In the first step, the reduced form $y_{2i} = X_i\alpha_1 + Z_i\alpha_2 + v_2$ is estimated by OLS. In the second step, the parameters of the exponential outcome model can be consistently estimated by PQML with the first-stage residuals $\hat{v}_2$ included as an additional regressor.[15]

7.3 Sample Selection

In the example of wage decompositions, there are often concerns that group 0 and 1 select differently into the labor market. This is innocuous as long as selection is explained by differences in X_i, but if selection is due to unobservables, some type of correction procedure is required.

In the log-linear model, a popular choice is the two-step approach due to Heckman (1976), which requires a joint normality assumption between the errors in the selection

[15] Technically, it is conceivable to use this IV method in combination with WPQML as well, but pursuing this further is beyond the scope of this paper.

model and the outcome model. Terza (1998) shows how this two-step approach can be extended to the exponential outcome model. The weakness of this procedure is the same as in the linear model; the coefficient estimates are inconsistent if the joint normality assumption is violated. In addition, credible identification usually requires an exclusion restriction.

Less restrictive methods can be used when panel data are available and confounding unobservables are thought to be time-invariant. For individuals where outcomes are observed at least once, missing outcomes in the other periods can be imputed. For example, Melly and Santangelo (2013) assume that individuals' position in the conditional outcome distribution is time-invariant. The advantage of such nonparametric imputation procedures is that distributional assumptions on outcome and selection models are not required.

8 Conclusions

This paper has argued that the standard Oaxaca-Blinder decomposition (Oaxaca, 1973; Blinder, 1973) based on logarithms brings about statistical problems because it is invalid if the data includes zero and because the log differential is not invariant to changes in higher-order moments of the distribution. Furthermore, the approach also raises conceptual issues because the log differential is an approximate percentage difference in geometric means, which is not an intuitive quantity and thus difficult to interpret.

We have therefore suggested modelling the original dependent variable directly to base the analysis on differences in arithmetic means. At the core of our study, we have proposed a new doubly-robust estimation strategy based on a propensity-score weighted Poisson quasi-maximum-likelihood (WPQML) estimator. This flexible parametric framework is suitable in many contexts of nonnegative outcomes and easy to implement in practice. Our Monte Carlo study has shown that the proposed estimator performs well in many circumstances when compared to a range of competing estimators. The WPQML estimator produces good results even under a misspecified outcome model or missing overlap in covariate distributions.

In our empirical application to the union wage gap in the United States, we find that quite different conclusions are reached depending on the decomposition approach taken. The arithmetic mean decomposition suggests that the union wage premium is much less important in explaining the union wage gap than the geometric mean decomposition, which is usually used in empirical studies. This finding is also reported in the analysis of the native-immigrant wage gap in Switzerland. We conclude that the measure of the mean can have important implications for decomposition analysis and deserves more careful consideration than it usually receives in empirical research.

References

Barsky, Robert, John Bound, Kerwin Ko' Charles, and Joseph P Lupton, "Accounting for the Black-White Wealth Gap," *Journal of the American Statistical Association*, 2002, *97* (459), 663–673.

Bauer, Thomas and Mathias Sinning, "An extension of the Blinder–Oaxaca decomposition to nonlinear models," *AStA Advances in Statistical Analysis*, 2008, *92* (2), 197–206.

_ , Silja Göhlmann, and Mathias Sinning, "Gender differences in smoking behavior," *Health Economics*, 2007, *16* (9), 895–909.

Black, Dan, Ameila Haviland, Seth Sanders, and Lowell Taylor, "Why do minority men earn less? A study of wage differentials among the highly educated," *The Review of Economics and Statistics*, 2006, *88* (2), 300–313.

Blackburn, McKinley L., "Estimating wage differentials without logarithms," *Labour Economics*, 2007, *14* (1), 73 – 98.

_ , "Are Union Wage Differentials in the United States Falling?," *Industrial Relations: A Journal of Economy and Society*, 2008, *47* (3), 390–418.

Blinder, Alan S., "Wage discrimination: reduced form and structural estimates," *Journal of Human Resources*, 1973, *8* (4), 436–455.

Bustamante, Arturo Vargas and Jie Chen, "Health expenditure dynamics and years of US residence: analyzing spending disparities among Latinos by citizenship/nativity status," *Health Services Research*, 2011, *47* (2), 794–818.

Cotton, Jeremiah, "On the decomposition of wage differentials," *The Review of Economics and Statistics*, 1988, *70* (2), 236–243.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik, "Dealing with limited overlap in estimation of average treatment effects," *Biometrika*, 2009, *96* (1), 187–199.

Darity, William, David Guilkey, and William Winfrey, "Ethnicity, race, and earnings," *Economics Letters*, 1995, *47* (3), 401–408.

Edin, Per-Anders and Katarina Richardson, "Swimming with the tide: Solidary wage policy and the gender earnings gap," *The Scandinavian Journal of Economics*, 2002, *104* (1), 49–67.

Fairlie, Robert W, "An extension of the Blinder-Oaxaca decomposition technique to logit and probit models," *Journal of Economic and Social Measurement*, 2005, *30* (4), 305–316.

Firpo, Sergio, Nicole Fortin, and Thomas Lemieux, "Decomposing Wage Distributions using Recentered Influence Function Regressions," 2007. University of British Columbia.

_ , _ , and _ , "Decomposition methods in economics," *Handbook of Labor Economics*, 2011, *4*, 1–102.

Gourieroux, Christian, Alain Monfort, and Alain Trognon, "Pseudo Maximum Likelihood Methods: Theory," *Econometrica*, 1984, *52* (3), 681–700.

Heckman, James J., "The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models," in "Annals of Economic and Social Measurement," Vol. 5, NBER, 1976, pp. 475–492.

_ and Solomon Polachek, "Empirical evidence on the functional form of the earnings-schooling relationship," *Journal of the American Statistical Association*, 1974, *69* (346), 350–354.

Imbens, Guido W., "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *The Review of Economics and Statistics*, 2004, *86* (1), 4–29.

_ and Jeffrey M. Wooldridge, "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 2009, *47* (1), 5–86.

Jann, Ben, "The Blinder-Oaxaca decomposition for linear regression models," *Stata Journal*, 2008, *8* (4), 453–479.

Jürges, Hendrik, "The distribution of the German public–private wage gap," *Labour*, 2002, *16* (2), 347–381.

Kaiser, Boris, "Detailed Decomposition in Generalized Linear Models.," 2013. Working Paper, University of Bern.

Kim, Chang H., "Decomposing the Change in the Wage Gap Between White and Black Men Over Time, 1980-2005: An Extension of the Blinder-Oaxaca Decomposition Method," *Sociological Methods & Research*, 2010, *38* (4), 619–651.

Krieg, John M and Paul Storer, "How much do students matter? Applying the Oaxaca decomposition to explain determinants of adequate yearly progress," *Contemporary Economic Policy*, 2006, *24* (4), 563–581.

Leslie, Derek and Phil Murphy, "Measuring discrimination by decomposing earnings functions," *Applied Economics Letters*, 1997, *4* (2), 117–120.

Machado, José A.F. and José Mata, "Counterfactual decomposition of changes in wage distributions using quantile regression," *Journal of Applied Econometrics*, 2005, *20* (4), 445–465.

Manning, Willard G. and John Mullahy, "Estimating log models: to transform or not to transform?," *Journal of Health Economics*, 2001, *20* (4), 461–494.

Melly, Blaise, "Decomposition of differences in distribution using quantile regression," *Labour Economics*, 2005, *12* (4), 577–590.

_ **and Giulia Santangelo**, "The evolution of the gender wage gap: 1968-2008," 2013. Working Paper.

Mincer, Jacob A., "Schooling, Experience, and Earnings," *NBER Books*, 1974.

Mueller, Steffen, "Works councils and establishment productivity," *Industrial and Labor Relations Review*, 2012, *65*, 880–975.

Mullahy, John, "Econometric modeling of health care costs and expenditures: a survey of analytical issues and related policy considerations," *Medical care*, 2009, *47*, S104–S108.

Munn, Ian A and Anwar Hussain, "Factors determining differences in local hunting lease rates: insights from Blinder-Oaxaca decomposition," *Land Economics*, 2010, *86* (1), 66–78.

Neumark, David, "Employers' discriminatory behavior and the estimation of wage discrimination," *Journal of Human Resources*, 1988, *23* (3), 279–295.

Oaxaca, Ronald L., "Male-female wage differentials in urban labor markets," *International Economic Review*, 1973, *14* (3), 693–709.

_ **and Michael R. Ransom**, "Identification in detailed wage decompositions," *Review of Economics and Statistics*, 1999, *81* (1), 154–157.

Robins, James, Mariela Sued, Quanhong Lei-Gomez, and Andrea Rotnitzky, "Comment: Performance of Double-Robust Estimators When Inverse Probability Weights Are Highly Variable," *Statistical Science*, 2007, *22* (4), 544–559.

Rothe, Christoph, "Decomposing the composition effect," 2012. Working Paper, Columbia University.

Rubin, Donald B, "Estimating causal effects of treatments in randomized and nonrandomized studies.," *Journal of Educational Psychology*, 1974, *66* (5), 688–701.

Santos Silva, João M.C.S. and Silvana Tenreyro, "The log of gravity," *The Review of Economics and Statistics*, 2006, *88* (4), 641–658.

Słoczyński, Tymon, "New Evidence on Linear Regression and Treatment Effect Heterogeneity," 2012. University of Warsaw.

Terza, Joseph V, "Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects," *Journal of Econometrics*, 1998, *84* (1), 129–154.

Wooldridge, Jeffrey M., "Some Alternatives to the Box-Cox Regression Model," *International Economic Review*, 1992, *33* (4), 935–955.

_ , "Inverse probability weighted estimation for general missing data problems," *Journal of Econometrics*, 2007, *141* (2), 1281–1301.

_ , *Econometric Analysis of Cross Section and Panel Data*, The MIT Press, 2010.

Yun, Myeong-Su, "Decomposing differences in the first moment," *Economics letters*, 2004, *82* (2), 275–280.

_ , "A simple solution to the identification problem in detailed wage decompositions," *Economic Inquiry*, 2005, *43* (4), 766–772.

Appendix

A Proof of Proposition 1

Part I. Consistency. The uniform law of large numbers implies $\hat{\beta}^0_{WPQML} \xrightarrow{p} \beta^{0*}$ such that $\frac{1}{N_1}\sum_{i:D_i=1}^{N_1} \exp(X_i\hat{\beta}^0_{WPQML}) \xrightarrow{p} E[\exp(X_i\beta^{0*})|D_i = 1]$.

Part II: Assume the CEF of the outcome model implied by WPQML is correct, i.e. $E[y_i|X_i, D_i = 0] = \exp(X_i\beta^0)\ \forall X_i \in \mathcal{X}$, where $\beta^0 \in \mathbb{R}^k$ is the vector of "true" coefficients. By applying the law of iterated expectations to the first-order conditions in (10), it follows immediately that $\beta^{0*} = \beta^0$. Assumptions (i)-(iii) then imply $E[y_i|X_i, D = 0] = E[y_{i0}|X_i, D = 0] = E[y_{i0}|X_i, D_i = 1] = \exp(X_i\beta^*)\ \forall X_i \in \mathcal{X}$. By the law of iterated expectations, we have $E[y_{i0}|D_i = 1] = E[\exp(X_i\beta^{0*})|D_i = 1]$.

Part III: Assume the model for the propensity score is correct in the sense that $plim[\hat{\Lambda}(X_i)] = p(X_i)$. Since X_i includes a constant the first order-conditions in (10) imply $E[\omega(X_i)(E[y_i|X_i, D_i = 0] - \exp(X_i\beta^{0*}))|D_i = 0] = 0$. Manipulating this expression, we obtain

$$\begin{aligned} E[\omega(X_i)E[y_{i0}|X_i, D_i = 0]|D_i = 0] &= E[\omega(X_i)\exp(X_i\beta^{0*})|D_i = 0] \\ \int E[y_{i0}|X_i = z]\omega(z)dF_{X|D=0}(z) &= \int \exp(z\beta^{0*})\omega(z)dF_{X|D=0}(z) \\ \int E[y_{i0}|X_i = z]dF_{X|D=1}(z) &= \int \exp(z\beta^{0*})dF_{X|D=1}(z) \\ E[y_{i0}|D_i = 1] &= E[\exp(X_i\beta^{0*})|D_i = 1] \end{aligned} \tag{12}$$

On the LHS, the second line follows from assumption (iii). On both sides, the third line follows from assumption (iii) and Bayes' rule which implies that $\omega(X_i) = \frac{p(X_i)}{1-p(X_i)}\frac{1-p}{p} = \frac{dF_{x|D=1}(X_i)}{dF_{x|D=0}(X_i)}$. The fourth line follows from the law of iterated expectations. ■

Note that for Part II of the proof, the weighting function is irrelevant because the same result follows for the unweighted PQML estimator.

B Monte Carlo Simulation

In the case of perfect common support, we set $\delta_1 = \delta_2 = 0.2$ such that the covariate distributions are very similar across groups. If we estimate a logit model for the propensity score, $p(X_i)$, the predicted values all lie between 0.1 and 0.9. Hence there is perfect common support. (We appeal to these particular threshold values because the findings of Crump et al. (2009) suggest that discarding observations outside this interval is a good approximation to the optimal cut-off rule.) In the second specification, we set $\delta_1 = \delta_2 = 0.85$ to induce large differences in the group-specific covariate-distributions. The values are chosen such that roughly 20% of the predicted propensity score values (as estimated from a logit model) are outside the interval $[0.1, 0.9]$. Hence, this represents a scenario in which the common-support assumption is severely violated. For illustration, Figure 4 shows kernel densities of the estimated propensity scores for the two assignment rules based on a random Monte Carlo sample. Comparing counterfactual estimates from these two assignment rules will allow us to assess the impact of the common-support assumption on the performance of the estimators.

The CEFs in the simulation exercise are parameterized as shown in Table A.1 below. The slope coefficients in β^0 are larger than the coefficients in β^1 to induce a structural effect between group 0 and group 1. The coefficients in the quadratic CEF model are larger than in the exponential CEF model because the curvature in the former is less pronounced. The coefficients are chosen such that mean outcomes are of similar size across the two specifications.

Figure 4: Kernel Densities of Estimated Propensity Scores

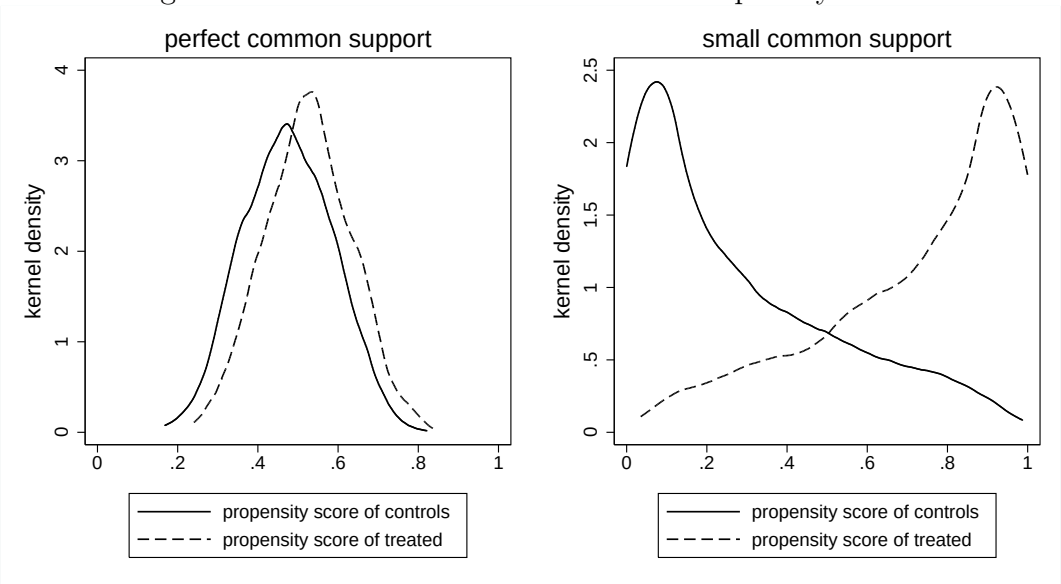


Table A.1: Parameterization for Monte Carlo

	exponential CEF		quadratic CEF	
coefficients	β^1	β^0	β^1	β^0
constant	5	5	23	23
X_1	0.3	0.36	3.5	4.55
X_2	0.2	0.24	3.5	4.55

In the homoskedastic case, we set $V[\varepsilon_g|x] = V[\varepsilon_g] = 0.5$, which implies $V[y_g|x] \propto E[y_g|x]^2$. Note that this is equivalent to assuming homoskedasticity in the log-linear regression model (Santos Silva and Tenreyro, 2006). In the heteroskedastic case, the error variance is proportional to the CEF, $V[\varepsilon_g|x] \propto E[y_g|x]$, which implies $V[y_g|x] \propto E[y_g|x]^3$. In the latter case, we scale the distribution of the error term such that $V[\varepsilon_g] = 0.5$. This enables us to analyse the impact of heteroskedasticity while holding the overall dispersion of the error distribution fixed.